Precipitation Missing Data Prediction Using Recommendation System

Herdianti Darwis, Fitiyani Umar

Abstract: Complete data is generally required in data analysis especially in time-series-related study. However, incomplete data due to equipment malfunction, human error, disaster, or other unknown reason is practically discovered. It is required to perform missing data prediction before forecasting the future values. Recommendation system is a system that predicts the "rating" or "preference" of a user over an item. Instead of dealing to a function of time series, the weekly precipitation data of Makassar City is placed into a matrix form consisting of "years" in row as the users and "weeks of the year" in column as the items. This method is also known as matrix decomposition. Accuracy of prediction by root mean square error (RMSE) and mean absolute error (MAE) have been performed to compare the predicted result by using the matrix decomposition to the observed values. In this study, matrix decomposition is discovered as a reliable method in dealing with the missing values of historical observation and forecasting the future values simultaneously.

Index Terms: Missing data, prediction; forecasting; recommendation system; matrix decomposition; RMSE; MAE.

1. INTRODUCTION

Complete data is generally required in data analysis especially in time-series-related study. However, incomplete data due to equipment malfunction, human error, disaster, or other unknown reason is practically discovered. It is required to perform missing data prediction before forecasting the future value [1], [2]. Research on handling missing value has been also performed in several approaches, such as Listwise Deletion, Pairwise Deletion, Mean Imputation, Regression Imputation, K-Means Imputation (KMI), Fuzzy K-Means clustering Imputation (FKMI), Support Vector Machine Imputation (SVMI) [3], [4]. On the other hand, ARIMA and k-nearest neighbor Neural Network (KNN) are two common methods used for forecasting in various fields including hydrology [5], engineering etc. [6]–[8]. However, the algorithms are much longer in the process of calculation when a prediction must be calculated quickly in real time. Recommendation system is a system that predicts the "rating" or "preference" of a user over an item. Recently, recommendation systems have become widely applied in a various applications, not only for rating movies, filtering books, or articles for reading lover, and recommending Twitter followers but also in traffic accident [9], [10]. In recommendation systems, input data is placed into an incomplete utility matrix; one dimension representing users and the other representing items of interest corresponding to row and column respectively. The blank elements are then predicted using the observed values. The method is known as matrix decomposition method which is first used in predicting of infectious disease spread and rainfall in 2013 [11], [12]. Precipitation is the condensation product of atmospheric water vapor falling under gravity which is commonly analyzed using forecasting. Indonesia lies along the equator causing a tropical climate so that precipitation is about rainfall. Forecasting future values of precipitation is supposed to be crucial in order to anticipate flood occurring within rainy season and water scarcity within dry season. However, it is undeniable that the existence of missing data could put us in trouble. Weekly precipitation data (January 1991- June 2017) taken from bureau of meteorology and climatology, Makassar, Indonesia with 109 of 1404 weeks are missing. Instead of handling missing values of the historical observation and forecasting the future values separately, this study shows time series data regarded into a matrix form as recommendation systems in dealing with missing values and future values prediction simultaneously within one process in order to optimize the result and minimize time consumption.

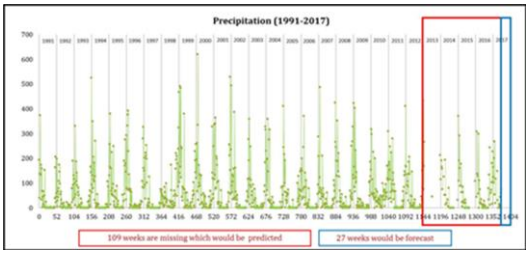

Fig. 1: *Plot of Makassar precipitation data, Indonesia from 1991-2017*

.

2 EXPERIMENTAL DETAIL

In this study, we have performed two experiments to the precipitation data with different range in time. The first experiment is to check the validity of the method, and the second one is handling the real missing values. Table 1 shows that the dataset is divided into two sets, a training data and a testing data with two kinds of case called case2012 and case2017

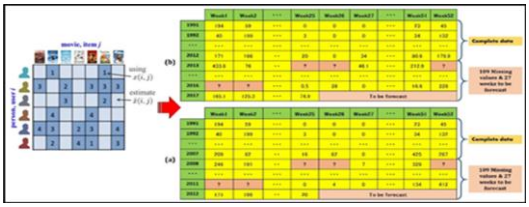

Fig. 2: *Utility matrices formation based on the main idea of the recommendation systems using the matrix decomposition*

- *Herdianti Darwis is currently lecturer in compuer science in University of Muslim Makassar, Indonesia. E-mail: : herdianti.darwis@umi.ac.id*
- *Fitiyani Umar is currently lecturer in compuer science in University of Muslim Makassar, Indonesia*

TABLE 1
TABEL TRAINING AND TESTING DATA

	Training Data	Testing Data
Case2012	1035 weeks (1991-2012)	109 weeks (2008-2012)
Case2017	1295 weeks (1991-	109 weeks (2013-

$$S = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}(A_{i,j} - \sum_{l=1}^{k} \boldsymbol{U_{i,l}^{T}}.\boldsymbol{V_{l,j}}) \quad + \quad \frac{\lambda_U}{2}\|U\|_{Fro}^2 + \frac{\lambda_V}{2}\|V\|_{Fro}^2 \quad (1)$$

As pointed in (1), λ_u and λ_v are regularization coefficients to prevent over fitting; $\|.\|_{Fro}^2$ denotes the Frebenius norm (l^2 norm). The optimization of U and V is performed by the descent gradient method or stochastic descent gradient method using the algorithm (2)

$$U^{t+1} = U^t - \mu\frac{\partial S}{\partial U} \text{ and } V^{t+1} = V^t - \mu\frac{\partial S}{\partial V} \quad (2)$$

As pointed in (2), μ is the leaning rate and λ is the regularization parameter. The root mean square error (RMSE) and mean absolute error (MAE) are used to measure the differences between value predicted by matrix decomposition and the values actually observed. As the name suggested, RMSE and MAE are defined (3)

$$RMSE = \sqrt{\frac{1}{|T|}\sum_{i,j} I(i,j)\left(\hat{A}(i,j) - A(i,j)\right)^2} \quad \text{and} \quad MAE = \sqrt{\frac{1}{|T|}\sum_{i,j} I(i,j)\left|\hat{A}(i,j) - A(i,j)\right|} \quad \text{Respectively}$$

$$\text{with } |T| = \sum_{i,j} I(i,j)\,,\ I(i,j)\text{: test} \quad (3)$$

3 RESULT AND DISCUSION

In the process, after trying for some learning rates, we discovered that the learning rate $\mu = 1e-5$ is the best for the data. We used symmetric regularization parameter $\lambda = 1e-8$ for both U and V and performed 10 kinds of dimension to find the optimized $U_i$and V_j. Table 2 shows the prediction accuracy of case2012 using RMSE and MAE i.e. missing data prediction, future value forecasting, and predicting and forecasting result simultaneously. MD (f=2) is chosen as the superior dimension to be used for case 2017 due to the smaller RMSE and MAE in both missing data prediction and future values forecasting of case2012.

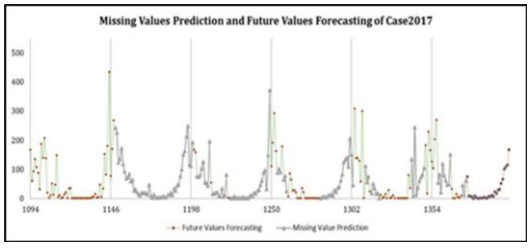

Fig. 3: *Plot of the predicted missing data and forecast future values of case 2017*

4 CONCLUSION

In conclusion, this study shows that handling missing values and forecasting the future values could be performed in one way using matrix decomposition in recommendation system. Matrix decomposition provided excellent results and fulfilled the top 5 rated small RMSE and MAE in case2012. Considering its success in case2012, we have performed the real missing data prediction in the last 5 years which is called case2017. Matrix decomposition is reliable and simple to be applied and could be one recommended method used in precipitation missing data prediction. In the future, matrix decomposition related to Bayesian and Markov process to optimize the U and V in order to increase the accuracy and minimize error.

TABLE 2
THE PREDICTION ACCURACY OF CASE2012 USING RMSE AND MAE VALUES FOR VALIDATION

Methods	Predicting Missing Value		Forecast Future Value		Simultaneously Calculation	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
MD (f = 1)	59.66 #3	35.95 #2	23.82 #3	14.49 #2	54.45 #3	31.69 #3
MD (f = 2)	59.44 #2	35.93 #1	23.88	14.56 #3	54.27 #2	31.68 #2
MD (f = 3)	61.78	38.21	23.22 #1	14.68	56.27	33.54
MD (f = 4)	77.64	46.67	31.53	20.43	70.91	41.47
MD (f = 5)	78.41	48.75	33.84	20.89	71.80	43.22
MD (f = 6)	63.13	40.05	31.46	17.19	58.23	35.51
MD (f = 7)	66.80	42.88	32.54	21.74	61.54	38.68
MD (f = 8)	72.58	45.68	23.41 #2	13.49 #1	65.81	39.29
MD (f = 9)	57.95 #1	38.04 #3	32.39	20.45	53.85 #1	34.55 #1
MD (f = 10)	68.82	43.75	40.62	25.89	64.22	40.21

ACKNOWLEDGMENT

The authors would like to thank Prof. Hideo Hirose for the guidance in understanding recommendation system. We also appreciate Universitas Muslim Indonesia for providing financial support.

REFERENCES

[1] S. Keliwar, A. Bramanto Wicaksono Putra, J. Hammad, and Haviluddin, "Modeling of time series data for forecasting the number of foreign tourists in east Kalimantan using fuzzy inference system based on ARX model," Int. J. Eng. Technol., 2018.

[2] R. Rofiqoh, A. F. O. Gaffar, D. Setyadi, and S. Hudayah, "Net income prediction of several leading bank in Indonesia using neural approach," Int. J. Eng. Technol., vol. 7, no. 2.2, pp. 99–103, 2018.

[3] R. L. Vaishnav and K. M. Patel, "Analysis of various techniques to handling missing value in dataset," Int. J. Innov. Emerg. Res. Eng., vol. 2, no. 2, pp. 191–195, 2015.

[4] Suprihatin, I. T. R. Yanto, N. Irsalinda, T. Purwaningsih, Haviluddin, and A. P. Wibawa, "A performance of

modified fuzzy C-means (FCM) and chicken swarm optimization (CSO)," in Proceeding - 2017 3rd International Conference on Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era, ICSITech 2017, 2018.

[5] Mislan, A. F. O. Gaffar, Haviluddin, and N. Puspitasari, "Water Level Prediction of Lake Cascade Mahakam Using Adaptive Neural Network Backpropagation (ANNBP)," in 1st International Conference on Tropical Studies and Its Application (ICTROPS), 2018.

[6] R. Y. M. Li, S. Fong, and K. W. S. Chong, "Forecasting the REITs and stock indices: group method of data handling neural network approach," Pacific Rim Prop. Res. J., vol. 23, no. 2, pp. 123–160, 2017.

[7] Y. Rong, X. Zhang, X. Feng, T. K. Ho, W. Wei, and D. Xu, "Comparative analysis for traffic flow forecasting models with real-life data in Beijing," Adv. Mech. Eng., vol. 7, no. 12, 2015.

[8] Purnawansyah and Haviluddin, "K-Means clustering implementation in network traffic activities," in Proceedings - CYBERNETICSCOM 2016: International Conference on Computational Intelligence and Cybernetics, 2017.

[9] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh, "Wtf: The who to follow service at twitter," in Proceedings of the 22nd international conference on World Wide Web, 2013, pp. 505–514.

[10] L. Barba and N. Rodríguez, "Traffic accidents forecasting using singular value decomposition and an autoregressive neural network based on PSO," Polibits, vol. 51, pp. 33–38, 2015.

[11] H. Hirose, T. Nakazono, M. Tokunaga, T. Sakumura, S. Sumi, and J. Sulaiman, "Seasonal infectious disease spread prediction using matrix decomposition method," in 2013 4th International Conference on Intelligent Systems, Modelling and Simulation, 2013, pp. 121–126.

[12] H. Hirose, M. Tokunaga, T. Sakumura, J. Sulaiman, and H. Darwis, "Matrix approach for the seasonal infectious disease spread prediction," IEICE Trans. Fundam. Electron. Commun. Comput. Sci., vol. 98, no. 10, pp. 2010–2017, 2015.